MVA2015 IAPR International Conference on Machine Vision Applications, May 18-22, 2015, Tokyo, JAPAN

The method based on view-directional consistency constraints for robust 3D object recognition

Jun SHIMAMURA, Taiga YOSHIDA, Yukinobu TANIGUCHI,
Hiroko YABUSHITA*, Kyoko SUDO, and Kazuhiko MURASAKI
NTT Media Intelligence Laboratories, NTT Corporation
1-1 Hikarinooka Yokosuka, Kanagawa, 239-0847, Japan
shimamura.jun@lab.ntt.co.jp

Abstract

This paper proposes a novel geometric verification method to handle 3D viewpoint changes under cluttered scenes for robust object recognition. Since previous voting-based verification approaches, which enable recognition in cluttered scenes, are based on 2D affine transformation, verification accuracy is significantly degraded when viewpoint changes occur for 3D objects that abound in real-world scenes. The method based on view-directional consistency constraints requires that the angle in 3D between observed directions of all matched feature points on two given images must be consistent with the relative pose between the two cameras, whereas the conventional methods consider the consistency of the spatial layout in 2D of feature points in the image. To achieve this, we first embed observed 3D angle parameters into local features when extracting the features. At the verification stage after local feature matching, a voting-based approach identifies the clusters of matches that agree on relative camera pose in advance of full geometric verification. Experimental results demonstrating the superior performance of the proposed method are shown.

1 Introduction

The goal of our work is object recognition in real-world scenes. That is, given a query image, we want to find similar images that contain the same objects in a database: a large corpus of images. This type of retrieval plays an important role in many applications, such as mobile visual search and object detection[15, 7, 10]. Most state-of-the-art object and image retrieval approaches [6, 14, 11, 18, 4, 8] adopt the standard Bag-Of-Words (BoW) model initially introduced in [17]. While this model generally works well, it suffers from a problem: the loss of spatial information when representing the images as histograms of quantized features.

To address the issue caused by BoW representation, robust regression-based geometric verifications are applied to eliminate false matches of local features [15, 7, 10, 12, 5, 16]. To verify the matches, the geometric transformation between the query and a candidate image selected from the reference images is usually estimated using robust regression techniques such as RANSAC or LMedS. The transformation is often represented by affine [12], homography, or epipolar geometry [15]. However, many well-known robust regression techniques perform poorly or calculation time increases when the percentage of outliers increases, as in cluttered scenes.

To alleviate the negative effect of false matches stemming from cluttered scenes, voting-based approaches using parameters of local features have been proposed [10, 5, 19]. A voting-based approach can extract feature groupings from cluttered images in linear time. In [10, 5, 7, 16], 2D affine transformation is estimated using the differences in four parameters of three local features: orientation, scale, and 2D location. Each pair of matches generates a set of parameters that vote in a 4D histogram. The differences in the parameters consistently have the same value when the matches are correct. These voting-based approaches are designed to handle object orientation, translation, and scaling in 2D space. In other words, they have strong assumptions and can only work appropriately when the entire image can be approximated by affine transforms in planar. If viewpoint changes take place, they are rendered useless by perspective projection distortion. In particular, if the subjects are 3D objects that abound in real-world scenes, verification accuracy is significantly degraded because corresponding points in two images do not follow affine transformation.

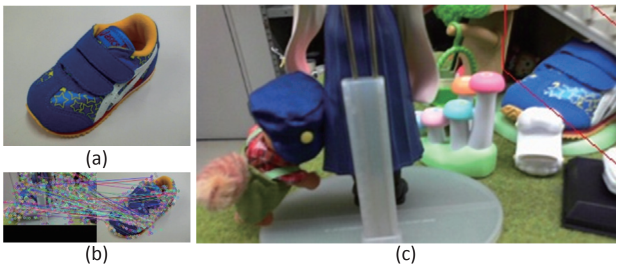
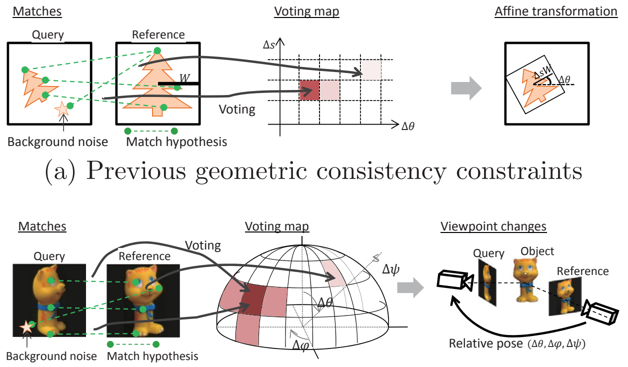


Figure 1. An example recognition result of our approach in a cluttered scene with a large 3D viewpoint change and extensive occlusion. (a) Reference image. (b) Result of local feature matching. (c) Recognition result. The recognized object lies within a red rectangle showing the boundaries of the original reference image.

In this paper, we propose a novel geometric verification method to handle 3D viewpoint changes under cluttered scenes. To overcome the consistency violation caused by viewpoint changes, we propose voting-based geometric verification using new constraints called view-directional consistency constraints. Such constraints require that the angle in 3D between observed directions of all matched feature points on two given images must be consistent with the relative pose between the two cameras, whereas the conventional methods consider the consistency of the spatial layout in 2D of feature points in the image. Our method can correctly verify the matches by considering 3D view directional changes in addition to 2D affine transformation. Moreover, our approach can robustly recognize and localize 3D objects in heavily cluttered and extensively occluded scenes like Fig. 1.

[*] Presently with Smart-life Planning Department, NTT docomo



(a) Previous geometric consistency constraints

(b) Our view-directional consistency constraints

Figure 2. Comparison of consistency constraints.

2 Proposed method

2.1 View-directional consistency constraints

The key idea of our method is to verify the consistency of relative camera pose for a set of matches of two given images. Fig. 2 illustrates a comparison of our method with previously reported schemes. As shown in Fig. 2(a), most previous voting-based approaches are based on consistent matches with an affine transformation between query and reference images. On the basis of differences in local feature parameters, each match votes for a correspondent grid of a voting map; this corresponds to an affine transformation. While correct matches form peaks in the voting map, failed matches stemming from the background scene may not form peaks. If the affine transformation assumption is broken, such as when a viewpoint change occurs with respect to the 3D object, even correct matches will no longer form the peaks in the voting map.

We focus on a relative pose between the view directions of a query camera and a reference camera. As shown in Fig. 2(b), the underlying idea is that if local features have observed angle parameters with three degrees of freedom (3DOF), the differences in the parameters for each match will be consistent for an object even when a view direction change occurs. We call this as view-directional consistency constraints. Two important things are needed for this idea to hold: (1) embedding observed angle parameters with 3DOF into local features; and (2) a voting scheme to find the best view directional changes. These two important things are described in the following subsections 2.2 and 2.3.

2.2 Embedding observed angle parameters with 3DOF into local features

To embed observed angle parameters with 3DOF into local features, we first assume that the 3D object is a combination of local patches that corresponds to the region of a local feature. Fig. 3(a) shows an affine transformed image of local patches on a 3D object with a view direction change. While the entire image cannot be approximated by affine transform, local patches in small regions of the image can be approximated by affine transformation, and these patches can match.

Incidentally, affine transformation $\mathbf{A}$ can be decomposed as a camera motion [12] as

$$\mathbf{A} = s\begin{bmatrix} cos\psi & -sin\psi \\ sin\psi & cos\psi \end{bmatrix}\begin{bmatrix} 1/cos\theta & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} cos\varphi & -sin\varphi \\ sin\varphi & cos\varphi \end{bmatrix}, \tag{1}$$

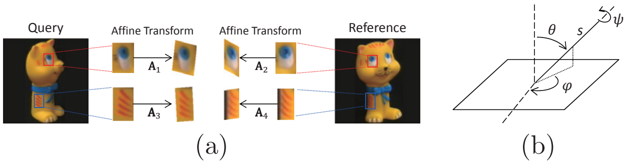


Figure 3. (a) Affine transformation on local patches, (b) Coordinate system

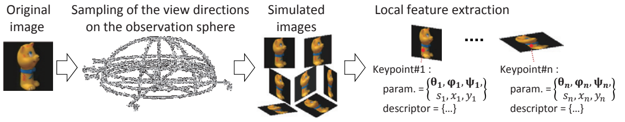

Original image
Sampling of the view directions on the observation sphere
Simulated images
Local feature extraction
Keypoint#1 :
param. = {θ1, φ1, ψ1, s1, x1, y1}
descriptor = {...}
Keypoint#n :
param. = {θn, φn, ψn, sn, xn, yn}
descriptor = {...}

Figure 4. Embedding observed angle parameters into local features.

where $s > 0$. Fig. 3(b) shows a camera motion interpretation of the affine decomposition: φ and θ are the viewpoint angles, ψ parameterizes the camera spin, and s is the camera zoom. This decomposition suggests that we can simulate view directional changes by varying the two camera axis parameters φ and θ, and generate affine transformed images. Therefore, we can embed observed angle parameters into local features extracted from each simulated image by appending the viewpoint angles φ and θ that were used to simulate view directional changes, as shown in Fig. 4. We can also embed the ψ and s parameters into local features by appending the orientation and scale parameters of local features respectively. Consequently, each local feature can be represented as $\mathbf{L} = \{\theta, \varphi, \psi, s, x, y\}$.

Like ASIFT [12], this simulation scheme allows us not only to estimate observed angle parameters to enable strict geometric verification, but also to extract affine invariant descriptors to enable matching under large pose changes. Since ASIFT is not a geometric verification method but an affine scale-invariant matching method, its recognition accuracy significantly increases when our geometric verification methods are applied to it to filter out false matches using the view-directional consistency constraints, as will be described later in Sec. 3.

2.3 Voting scheme to find possible view directional changes

After completing the local feature matching among all simulated images, the proposed method verifies the matches on the basis of the view-directional consistency constraints described in subsection 3.1.

Voting one match each to all relative camera poses would identify clusters of matches with a consistent interpretation. When clusters of matches are found to vote for the same relative pose with regard to a reference pose that captures a reference image, the probability of the interpretation being correct is much higher than that for any single match. Each of our local features specifies six parameters: 3DOF rotation (φ, θ, and ψ), scale (s), and 2D location (x, y). Therefore, we can create a voting map predicting the relative pose and scale from the match hypothesis. We do not use location parameters because the relative displacement of local features is not consistent on 3D objects. Consequently, our voting map is 4D as shown in Fig. 5. In similar fashion to [10], we use large grid sizes of 30 degrees for $\Delta\psi \in [0, 360)$, a scale factor of 2. The grid sizes for $\Delta\varphi \in [-180, 180]$ and $\Delta\theta \in [-90, 90]$ are determined experimentally (described in Sec. 3). In the

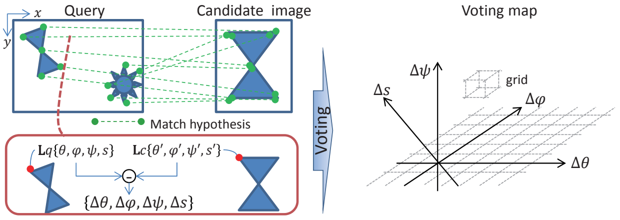

x
y
Query
Candidate image
Match hypothesis
$\mathbf{L}q\{\theta,\varphi,\psi,s\}$
$\mathbf{L}c\{\theta',\varphi',\psi',s'\}$
$\{\Delta\theta,\Delta\varphi,\Delta\psi,\Delta s\}$
Voting
Voting map
$\Delta\psi$
Δs
grid
$\Delta\varphi$
$\Delta\theta$

Figure 5. Voting map.

view direction sampling stage described in subsection 3.2, the sampling interval of the $\Delta\theta$ is more coarse than that of $\Delta\varphi$. Therefore, we use a larger grid size for $\Delta\theta$ than for $\Delta\varphi$. To avoid the problem of boundary effects in grid assignment, each match votes for the two closest grids.

2.4 Algorithm

The object retrieval algorithm that incorporates the proposed geometric verification method described in this subsection. The method first extracts local features and their descriptors from affine simulated query images with observed angle parameters embedding. Subsequently, local feature matching is performed between a query and a candidate image in the database. All local features $\mathbf{L_q}$ extracted from simulated query images are compared to all local features $\mathbf{L_c}$ extracted from simulated candidate images by their descriptors. Poor matches are rejected through a ratio test [10]. Then the proposed geometric verification based on view-directional consistency constraints is conducted. On the basis of differences in local feature parameters, each match votes for a correspondent grid of a voting map; this corresponds to a relative pose. All voters are equal as

$$V_c(\Delta\varphi,\Delta\theta,\Delta\psi,\Delta s) \leftarrow V_c(\Delta\varphi,\Delta\theta,\Delta\psi,\Delta s)+1, \quad (2)$$

where $V_c(.,.,.,.)$ means the number of votes in a correspondent grid, and $\leftarrow$ means the update procedure. This voting procedure identifies all clusters with at least seven entries, which is a sufficiently number to solve for epipolar geometry. To achieve stricter verification, each such cluster ($V_c(\Delta\varphi,\Delta\theta,\Delta\psi,\Delta s) \geq 7$) then undergoes a full geometric verification procedure as a hypothesis. We use the LO-PROSAC algorithm [3, 2] to estimate epipolar geometry, then discard outliers that slipped into the cluster due to large grid size or other errors. If fewer than seven matches remain after discarding the outliers, then the cluster is rejected. The final decision to accept or reject the hypothesis and object localization is performed on the basis of Lowe's probabilistic model given in [9]. Finally, the system outputs the total number of inliers for each candidate image as the score. The score S_c for the candidate image c is defined as

$$S_c = \sum^{N} g(\Delta\varphi,\Delta\theta,\Delta\psi,\Delta s), \quad (3)$$

where N means the total number of grids within the voting space, and g is a scoring function defined as

$$g(\Delta\varphi,\Delta\theta,\Delta\psi,\Delta s) = \begin{cases} p_i & \text{if the grid is accepted} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

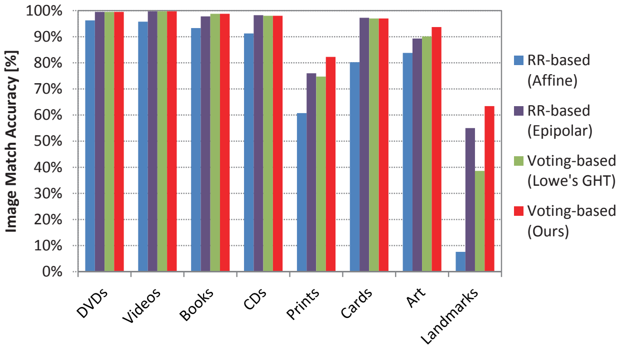


Figure 6. Accuracy comparisons.

Table 1. Comparison of mAP on the *INRIA Holidays* and *Oxford Building 5k* datasets.

	RR-based		Voting-based	
Datasets	Affine	Epipolar	Lowe's GHT	*Ours*
INRIA	0.487	0.794	0.754	**0.837**
Oxford5k	0.271	0.642	0.510	**0.706**

where p_i means the number of inliers, which forms a set of matches close to the estimated epipolar line, in a correspondent grid. For objects that project to small regions of an image, only seven matches may be sufficient to achieve reliable recognition.

3 Experimental results

3.1 Datasets and implementation details

We implemented the proposed method on a desktop PC with an Intel Core i7 3.33GHz processor running on Windows 7. We used SIFT [10] local features and FLANN [13] for matching local features. We evaluated our approach on three public datasets: *Stanford Mobile Visual Search* [1], *INRIA Holidays* [5], and *Oxford Building 5k* [15]. Every image was resized so that the longer side was 400 pixels in length.

In evaluating our approach, we measured the retrieval accuracy on the *Stanford Mobile Visual Search* dataset with image match accuracy as in [1]. As was done with most previous methods, the retrieval accuracy performance for the other datasets was measured with the mean average precision (mAP).

3.2 Evaluation results

We compared our method with three baseline geometric verification methods: (1) affine transform estimation with a robust regression technique and (2) epipolar geometry estimation with a robust regression technique as robust regression-based approaches (RR-based), and (3) Lowe's method [10] as a voting-based approach. To make a fair comparison of the performance by geometric verification, we used ASIFT as an affine scale-invariant matching method and LO-PROSAC as a robust regression technique in all four methods.

Fig. 6 shows accuracy comparisons of our method with the other baseline methods on *Stanford Mobile Visual Search*. It can be seen the proposed method outperforms the baselines in all categories. It is notewor-

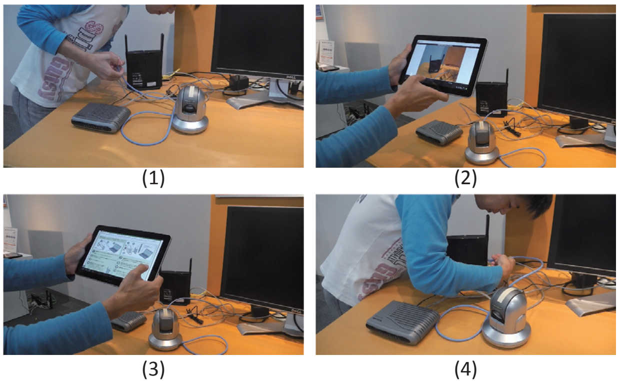


Figure 7. Example "Manual on the spot" application. (1) A user unfamiliar with the way to handle an appliance was having problems. Because many cables were connected already, it was difficult to turn the appliance over to find the model number. (2) The user got help from our system simply by using the camera. (3) The system recognized the appliance, enabling the user to browse the manual for it. (4) Finally, the user was able to handle the appliance properly.

thy that the accuracy of our method is much higher than that of the other baseline methods for the *Landmarks* category. This shows that the proposed view-directional consistency constraints work well even if viewpoint changes occur on 3D objects such as landmarks. With 3.33 GHz CPUs, the average matching and verification time per candidate image in the *Landmarks* category was 0.09s for the proposed method, as compared with 0.13s for baseline method (2) in second place This shows that the proposed method eliminated outliers more effectively than the robust regression techniques. Results obtained for other categories show that our method can correctly identify a reference image with no less accuracy than the baseline methods.

We also compared our method to the baseline methods on *INRIA Holidays* and *Oxford Building 5k*, as listed in Table 1. Here again we can see the proposed method outperforms the baseline methods.

3.3 Application

To confirm the feasibility of the proposed method, we implemented an application prototype we call "Manual on the spot". Fig. 7 shows an example scenario of a "Manual on the spot" application. When the user captures an appliance on the desk, the manual for the appliance is displayed on the tablet. As shown in the figure, the user got help from our system merely by picking up the camera in the living space, one in which there were many appliances.

In the application, we combined our method with a standard BoW-based retrieval method by re-ranking the shortlist of the retrieved images as the post-verification process. The robust 3D object recognition makes it possible to identify the appliances that the users are watching in real-world scenes.

4 Conclusion

We propose a novel method that introduces view-directional consistency constraints into geometric verification, making it capable of handling 3D viewpoint changes. In the proposed method, observed 3D angle parameters are embedded into local features. At the verification stage after feature matching, a voting-based approach identifies the clusters that agree on relative pose in advance of full geometric verification. The method achieves significantly improved verification and retrieval performance, especially for viewpoint changes with respect to the 3D objects that abound in real-world scenes. Our method can be combined with other BoW based retrieval methods to improve mean average precision (mAP) by re-ranking the list of the retrieved images as the post-verification process.

References

[1] V. R. Chandrasekhar et al. The stanford mobile visual search data set. In *ACM Multimedia Systems*, pages 117–122, 2011.

[2] O. Chum et al. Locally optimized RANSAC. In *DAGM*, pages 236–243, 2003.

[3] O. Chum and J. Matas. Matching with PROSAC - progressive sample consensus. In *CVPR*, pages 220–226, 2005.

[4] M. Jain et al. Asymmetric hamming embedding: taking the best of our bits for large scale image search. In *ACM MM*, pages 1441–1444, 2011.

[5] H. Jégou et al. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, pages 304–317, 2008.

[6] H. Jégou et al. On the burstiness of visual elements. In *CVPR*, pages 1169–1176, 2009.

[7] B. Leibe et al. Combined object categorization and segmentation with an implicit shape model. In *ECCV workshop on statistical learning in computer vision*, pages 17–32, 2004.

[8] Z. Liu et al. Embedding spatial context information into inverted file for large-scale image retrieval. In *ACM MM*, pages 199–208, 2012.

[9] D. G. Lowe. Local feature view clustering for 3d object recognition. In *CVPR*, pages 682–688, 2001.

[10] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, January 2004.

[11] A. Mikulik et al. Learning a fine vocabulary. In *ECCV*, pages 1–14, 2010.

[12] J. M. Morel and G. Yu. ASIFT: a new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, 2:436–469, April 2009.

[13] M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *VISSAPP*, pages 331–340. INSTICC Press, 2009.

[14] M. Perdoch et al. Efficient representation of local geometry for large scale object retrieval. In *CVPR*, pages 9–16, 2009.

[15] J. Philbin et al. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, pages 1–8, 2007.

[16] X. Shen et al. Object retrieval and localization with spatially-constrained similarity measure and k-NN reranking. In *CVPR*, pages 3013–3020, 2012.

[17] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477, 2003.

[18] X. Wang et al. Contextual weighting for vocabulary tree based image retrieval. In *ICCV*, pages 209–216, 2011.

[19] Y. Zhang et al. Image retrieval with geometry-preserving visual phrases. In *CVPR*, pages 809–816, 2011.